Cell PRESS
Structure
Article

Discovery of a Dipeptide Epimerase Enzymatic Function Guided by Homology Modeling and Virtual Screening

Chakrapani Kalyanaraman,[1,5] Heidi J. Imker,[3,5] Alexander A. Fedorov,[4] Elena V. Fedorov,[4] Margaret E. Glasner,[2] Patricia C. Babbitt,[1,2] Steven C. Almo,[4,*] John A. Gerlt,[4,*] and Matthew P. Jacobson[1,*]
[1]Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, 600 16th Street, San Francisco, CA 94158, USA
[2]Department of Biopharmaceutical Sciences, School of Pharmacy, University of California, 1700 4th Street, San Francisco, CA 94158, USA
[3]Departments of Biochemistry and Chemistry, University of Illinois at Urbana-Champaign, 600 South Mathews Avenue, Urbana, IL 61801, USA
[4]Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY 10461, USA
[5]These authors contributed equally to this work
*Correspondence: almo@aecom.yu.edu (S.C.A.), j-gerlt@uiuc.edu (J.A.G.), matt.jacobson@ucsf.edu (M.P.J.)
DOI 10.1016/j.str.2008.08.015

SUMMARY

We have developed a computational approach to aid the assignment of enzymatic function for uncharacterized proteins that uses homology modeling to predict the structure of the binding site and in silico docking to identify potential substrates. We apply this method to proteins in the functionally diverse enolase superfamily that are homologous to the characterized L-Ala-D/L-Glu epimerase from *Bacillus subtilis*. In particular, a protein from *Thermotoga martima* was predicted to have different substrate specificity, which suggests that it has a different, but as yet unknown, biological function. This prediction was experimentally confirmed, resulting in the assignment of epimerase activity for L-Ala-D/L-Phe, L-Ala-D/L-Tyr, and L-Ala-D/L-His, whereas the enzyme is annotated incorrectly in GenBank as muconate cycloisomerase. Subsequently, crystal structures of the enzyme were determined in complex with three substrates, showing close agreement with the computational models and revealing the structural basis for the observed substrate selectivity.

INTRODUCTION

Reliable assignment of function to proteins discovered in genome sequencing projects is a major challenge in genomic biology. Functional assignment of uncharacterized proteins is commonly accomplished by sequence analysis, but the assignment of function on the basis of homology can lead to incorrect or misleading annotations and cannot identify new functions. Assigning enzymatic function to proteins identified in genome sequencing efforts is challenging, in part because there is no simple relationship between measures of sequence similarity (e.g., sequence identity) and protein function. Highly similar proteins (60% sequence identity or greater) can catalyze distinct reactions, whereas highly divergent proteins can catalyze identical chemical transformations (Broun et al., 1998; de Souza et al., 1998; Seffernick et al., 2001; Tian and Skolnick, 2003). Misannotation of enzymatic function is severe in functionally diverse superfamilies, such as the enolase superfamily of $(\beta/\alpha)_8$ barrel enzymes considered here (Pegg et al., 2006; A. Schnoes and P.C.B., unpublished data). We and others have developed computational methods intended to help address this challenge. A central theme in several of these approaches is the use of structural as well as sequence information. One class of methods analyzes features of active sites (Barker and Thornton, 2003; Cammer et al., 2003; Polacco and Babbitt, 2006; Tremblay et al., 2006; Wangikar et al., 2003). Virtual metabolite screening against active sites has also been used successfully to identify enzyme substrates in retrospective (Favia et al., 2008; Hermann et al., 2006; Kalyanaraman et al., 2005; Macchiarulo et al., 2004) and prospective (Hermann et al., 2007; Song et al., 2007) studies. In general, this approach cannot be expected to predict the optimal substrate for an enzyme in terms of k_{cat}/K_M, in part because k_{cat} is extremely difficult to predict. However, as in the field of drug design, virtual screening can help to prioritize for experimental testing compounds that are more likely to bind to the active site, which is a prerequisite for catalytic activity.

Despite remarkable advances in structural biology, including the contributions of the Protein Structure Initiative (PSI), the number of protein sequences identified through genome sequencing continues to vastly outpace the rate of structure determination. Consequently, for the foreseeable future, structural models of most protein sequences will only be available by homology modeling approaches. In principle, sufficiently accurate computational methods would enable the construction of models that could be used as surrogates for experimentally determined structures, for example for drug discovery (Jacobson and Sali, 2004; Kenyon et al., 2006) or for understanding sequence-structure-function relationships, our focus here. In a previous study, we demonstrated that it was possible to predict the substrate specificity of a divergent member of the enolase superfamily encoded by *Bacillus cereus*, based on docking against a homology model (Song et al., 2007). Subsequent enzymatic characterization and crystallographic analysis confirmed the predictions.

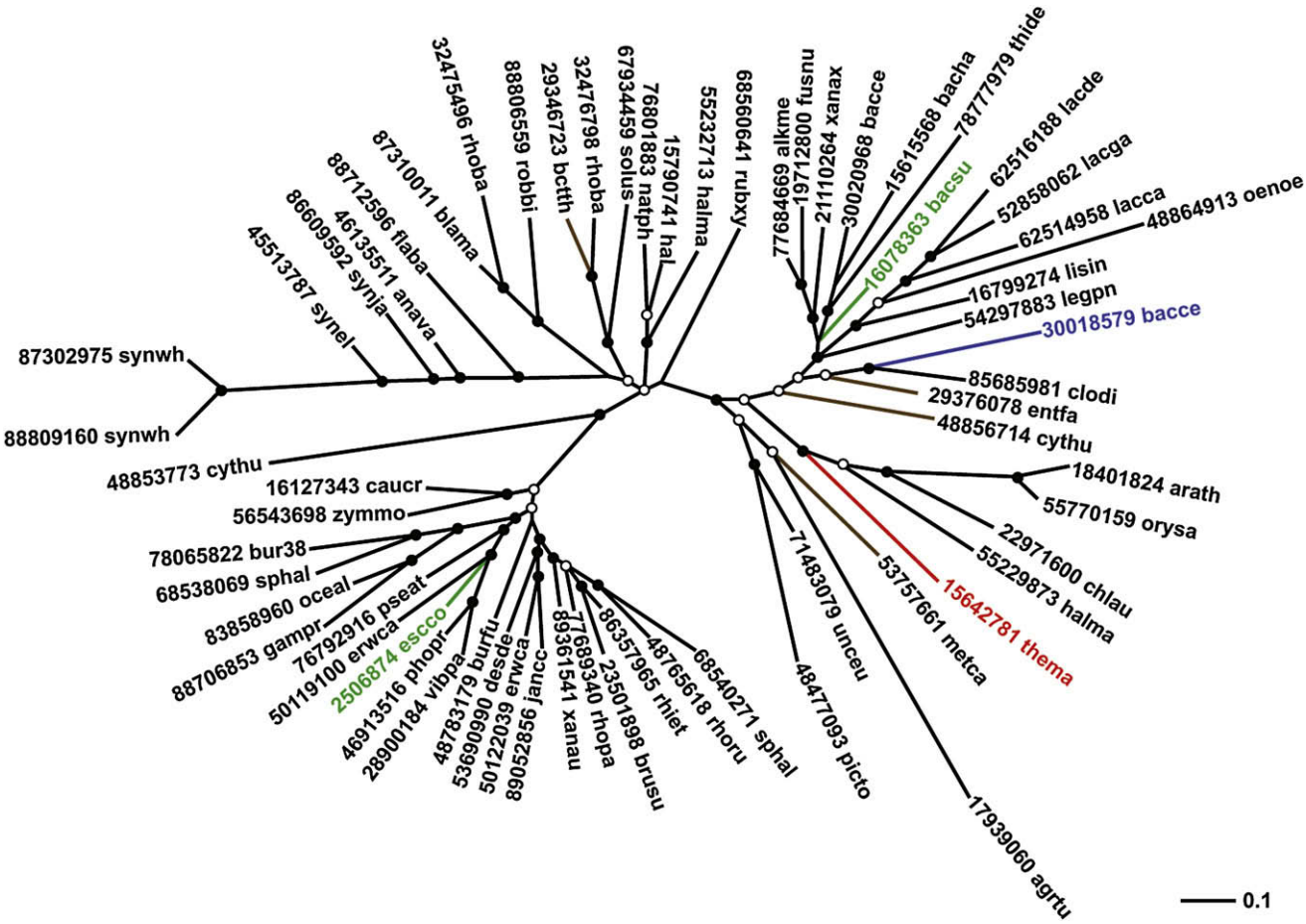

Figure 1. Phylogenetic Tree of a Representative Subset of the Dipeptide Epimerase Group of the Enolase Superfamily
The subset shown shares <60% pairwise sequence identity. *B. subtilis* and *E. coli* AEEs are green (Schmidt et al., 2001), the characterized *N*-succinyl-Arg/Lys racemase is blue (Song et al., 2007), and TM0006 is red.

Here, we have undertaken functional assignment for proteins in the enolase superfamily that are related most closely to the experimentally characterized L-Ala-D/L-Glu epimerases (AEEs) from *Escherichia coli* and *B. subtilis*, which are believed to be involved in recycling peptidoglycan (Klenchin et al., 2004). Although these proteins have a variety of annotations in GenBank, the most likely annotation based on careful phylogenetic analysis is AEE (Glasner et al., 2006). However, some evidence suggested that some of these proteins might have other functions; for example, this clade of proteins contains a few sequences from plants and archaea that lack peptidoglycan and hence have no obvious reason to encode the AEE function in their genomes. The purpose of this study was to identify sequences in this clade that might have alternative functions and to identify their substrates. In particular, we identify an enzyme with dipeptide epimerase activity in *Thermotoga maritima* with a novel specificity for dipeptides with alanine in the first position and aromatic amino acids in the epimerized position. Specifically, L-Ala-L-Phe, L-Ala-L-Tyr, and L-Ala-L-His are epimerized with values of $k_{cat}/K_M \sim 10^4\ M^{-1}\ s^{-1}$. Crystal structures of the enzyme have been determined in complex with three substrates, showing close agreement with the computationally predicted models and revealing the structural basis for the observed substrate selectivity.

RESULTS

Homology models were constructed for over 100 proteins in the MLE subgroup, including 82 sequences that clustered with the experimentally characterized *E. coli* and *B. subtilis* AEEs, by using the Protein Local Optimization Program from a multiple sequence alignment, as described in Experimental Procedures. Only 65 of these proteins are shown in the phylogenetic tree (Figure 1), which shows a representative subset sharing < 60% sequence identity. The template protein for all of the homology models was chosen to be 1TKK, the AEE from *B. subtilis* in complex with the L-Ala-L-Glu substrate. Apo structures are also available for both the *B. subtilis* and *E. coli* AEEs, but the binding sites are partially open, making them poorly suited to our purposes (Kalyanaraman et al., 2005).

In general, one challenge associated with metabolite virtual screening is that existing metabolite libraries are undoubtedly incomplete (for example, the specific dipeptides that are shown to be substrates here are not included in KEGG). We

Table 1. Top L/L Dipeptides from Docking against the Homology Models of TM0006, the *E. coli* AEE, and Four Other Representative Proteins, as Well as the Template Used for Those Models

	B. subtilis AEE (Crystal Structure)	*E. coli* AEE (Model)	48765618.rhoru (Model)	77684669.alkme (Model)	56543698.zymmo (Model)	67934459.solus (Model)	TM0006 (Model)
1	Ser-Glu	Thr-Glu	Asn-Glu	Cys-Glu	Lys-Glu	Ser-Glu	Ser-Trp
2	Cys-Glu	Ser-Glu	Thr-Glu	Ser-Glu	Glu-Glu	Cys-Asp	Thr-Trp
3	Thr-Glu	Gly-Glu	Ser-Glu	Asn-Glu	His-Glu	Ser-Asp	Met-Phe
4	Ser-His	Ser-His	Ile-Glu	Thr-His	Gln-Glu	Thr-Asn	Ser-Phe
5	Thr-His	Cys-Gln	Lys-Glu	Gly-Glu	Cys-Glu	Cys-Asn	Cys-Trp
6	Gly-Glu	Cys-His	Gln-Asp	Ala-Glu	Ser-Glu	Gly-Glu	Ile-Phe
7	Ser-Gln	Cys-Leu	Leu-Glu	Cys-His	Cys-Gln	Leu-Glu	Thr-Tyr
8	Ala-Glu	Cys-Met	Cys-Glu	Ser-His	Ser-Gln	Asn-Glu	Ile-Tyr
9	Gly-His	Asn-Glu	Gly-Glu	Thr-Glu	Thr-Glu	Cys-Glu	Met-His
10	Val-Glu	Val-Gln	Asn-Asp	Cys-Gln	Met-Gln	Thr-Leu	Ile-His

1TKK (AEE from *B. subtilis*) is the template used for the homology models.

hypothesized that all of the proteins considered in this study (Figure 1) were likely to be dipeptide epimerases, based on a phylogenetic tree of a larger subgroup in which the AEEs form a single clade (Glasner et al., 2006), and the conservation of catalytic residues and a DxD motif involved in binding the NH_3^+ terminus of the dipeptide in the AEEs. Accordingly, we restricted the virtual screening to the 400 possible L/L dipeptides. For computational efficiency, the protein was treated as rigid.

In control docking calculations with the *B. subtilis* AEE structure, L-Ala-L-Glu ranked 8 out of the 400 dipeptides; most of the other top-ranked dipeptides also had Glu or Asp at the epimerized position, and a small amino acid (Gly, Cys, Ser, Ala) in the first position (Table 1). Docking against a homology model of the *E. coli* AEE (32% sequence identity) led to similar results. It should be noted that both the *E. coli* and *B. subtilis* AEEs epimerize dipeptides other than L-Ala-L-Glu, which is believed to be the physiologically relevant substrate, albeit with slower kinetics. For example, both epimerize L-Ser-L-Glu and L-Ala-L-Met, and the *E. coli* AEE, which is the less specific of the two, epimerizes substrates such as L-Ala-L-His and L-Ala-L-Gln (Schmidt et al., 2001). Kinetic constants have been measured for only selected substrates, but suggest roughly 1 order of magnitude slower kinetics for nonphysiological substrates, i.e., k_{cat}/K_M for epimerizing L-Ala-D-Glu is 7.7×10^4 and 4.7×10^4 ($M^{-1}\,s^{-1}$) for the *E. coli* and *B. subtilis* AEEs, respectively, whereas the corresponding rates for L-Ala-D-Met are 2.8×10^3 and 2.2×10^3.

For most of the other homology models, especially those clustering relatively closely with the *E. coli* and *B. subtilis* AEEs, the docking results were similar to those obtained with the *E. coli* and *B. subtilis* AEEs, and thus consistent with the AEE activity. That is, the top hits were dominated by compounds with small amino acids in the first position, and negatively charged amino acids in the second, epimerized position. Four representative examples are shown in Table 1.

However, for ~20 of the proteins, the predicted specificities were dramatically different. Two major classes of novel predicted specificity were observed: a small number of enzymes (6) were predicted to epimerize positively charged dipeptides, and a somewhat larger number (~15) were predicted to epimerize hydrophobic (in both C- and N-terminal positions) dipeptides. Of these, we have obtained extensive experimental results (kinetics and multiple crystal structures) for the protein from *Thermotoga maritima* (gi:15642781, TM0006), confirming the computational predictions. Screening and structural studies are underway for several others, and those studies will be reported in due course.

The docking results for the homology model of TM0006, which shares 27% sequence identity with the *B. subtilis* AEE, are shown in Table 1. In the C-terminal, epimerized position, the docking results suggested selectivity for primarily aromatic, hydrophobic amino acids, instead of the strong selectivity for Glu in *B. subtilis* AEE. In the N-terminal position, top hits included Ser/Thr/Cys as well as larger hydrophobic amino acids such as Ile.

Experimental screening of L/L dipeptide libraries by mass spectroscopy (MS) confirmed the specificity switch (Table 2). In the Gly-Xxx, Ala-Xxx, and Thr-Xxx libraries, the best substrates had Phe, Tyr, or Trp in the epimerized position. Aliphatic side chains (Met, Leu, Ile) were also tolerated, and Ala-His and Thr-His were good substrates. In the N-terminal position, any hydrophobic amino acid was tolerated in the Xxx-Phe, Xxx-Tyr, and Xxx-His libraries. Dipeptides with charged, or most polar amino acids in the first position were usually poor substrates. Furthermore, the enzyme displayed no detectable muconate lactonizing enzyme (MLE) activity (results not shown), demonstrating that the GenBank and UniProt/TrEMBL annotations are incorrect. AEEs are part of a larger subgroup within the enolase superfamily, whose members are more similar to each other than other subgroups within the superfamily. The known functions of the subgroup are MLE, *o*-succinylbenzoate synthase, and racemization of *N*-succinyl or *N*-acetyl amino acids. No activity for these other assigned functions within the MLE subgroup was observed (data not shown).

Although MS screening of dipeptide libraries allowed us to simultaneously evaluate multiple substrates efficiently, we were only able to ascertain a rough approximation of activity. However, taking the MS screening results as a whole allowed us to prioritize our choice of substrates to carry out full kinetic assays. Kinetic constants were determined for selected dipeptide substrates by observing the change in optical rotation by

Table 2. Experimental Screening of TM0006 with L/L Dipeptides by Mass Spectroscopy to Detect Incorporation of Deuterium as a Result of Epimerization

	Varying C Terminus			Varying N Terminus			
	Gly-Xxx	Ala-Xxx	Thr-Xxx	Xxx-Phe	Xxx-Tyr	Xxx-His	Xxx-Lys
Ala				++	++	+++	++
Val	+			+	+	++	
Leu	+	++	+	++	++	++	+
Ile	+	++	+	++	++	++	+
Met	++	+	++	++	++	+++	+
Asp					+		
Glu	+						
Lys		+					
Arg		+	+	+			
Phe	++	+++	+++	+	++	+++	++
Trp	++	++	+	++	+	+++	+
Tyr	+++	+++	++	++	++	+++	+
Ser		+			++	+++	
Thr						+	
Asn						+	
Gln		+				+	
His	+	+++	+++		+		
Pro				+	+	+	

Deuterium incorporation was measured at 3 hr after incubation. "+," <25% incorporation; "++," 25%–50%; "+++," >50%. Empty cells in the table indicate no detectable incorporation. No incorporation was observed with any D-Ala-L-Xxx dipeptides, *N*-succinyl amino acids, *N*-acyl amino acids, or unmodified amino acids, which are thus omitted from the table.

Table 3. Kinetic Constants Obtained for Epimerization of Selected Dipeptide Substrates of TM0006

	k_{cat} (s^{-1})	K_M ($\times 10^{-3}$ M)	k_{cat}/K_M ($M^{-1}\ s^{-1}$)
L-Ala-L-Phe 28°C	16 ± 7.1	1.3 ± 0.56	(1.2 ± 0.20) × 10^4
L-Ala-L-Phe 40°C	35 ± 6.0	0.90 ± 0.26	(4.1 ± 0.87) × 10^4
L-Ala-L-Phe 50°C	76 ± 24	1.4 ± 0.31	(5.4 ± 0.42) × 10^4
L-Ala-L-Tyr	6.5 ± 0.31	0.71 ± 0.08	(9.1 ± 0.78) × 10^3
L-Ala-L-His	60 ± 4.9	5.3 ± 2.4	(1.3 ± 0.57) × 10^4
L-Ala-L-Glu	14 ± 4.2	2.8 ± 0.96	(4.9 ± 1.1) × 10^3
L-Ala-L-Leu	10 ± 0.72	2.9 ± 0.89	(3.8 ± 1.2) × 10^3
L-Ala-L-Lys	4.6 ± 1.7	13 ± 4.5	(3.6 ± 0.15) × 10^2
L-Phe-L-Phe	0.21 ± 0.031	0.33 ± 0.046	(6.3 ± 1.1) × 10^2
L-Thr-L-Phe	–	–	(1.1 ± 0.058) × 10^{3a}
L-His-L-Phe	0.19 ± 0.021	3.2 ± 0.27	(5.9 ± 0.83) × 10^1
L-Asp-L-Phe	n.d.[b]	n.d.	n.d.
L-Ile-L-Phe	8.4 ± 2.0	4.7 ± 1.4	(1.8 ± 0.17) × 10^3
L-Lys-L-Phe	10 ± 0.82	3.9 ± 0.68	(2.7 ± 0.57) × 10^3
L-Phe-L-His	–	–	(5.9 ± 3.0) × 10^{1c}
L-Ser-L-His	–	–	(3.6 ± 0.14) × 10^{3a}
L-Met-L-His	n.d.[b]	n.d.	n.d.
L-Lys-L-His	–	–	(3.3 ± 0.17) × 10^{3a}

Errors presented represent standard deviations from a minimum of three independent kinetic characterizations.
[a] Exhibits substrate inhibition at concentrations above 5 mM.
[b] Not detectable.
[c] Could not saturate.

polarimetry (Table 3). Using the *E. coli* and *B. subtilis* AEEs as standards, we expected that authentic substrates would exhibit values of k_{cat}/K_M in the $10^4\ M^{-1}\ s^{-1}$ range (Schmidt et al., 2001). Of the L-Ala-L-Xxx dipeptides assayed, L-Ala-L-Phe and L-Ala-L-His displayed values of 1.2 ± 0.2 × $10^4\ M^{-1}\ s^{-1}$ and 1.3 ± 0.6 × $10^4\ M^{-1}\ s^{-1}$, respectively. Although generally grouped with polar amino acids, histidine is also aromatic. Likewise, we found that L-Ala-L-Tyr was also epimerized with an appreciable efficiency of 9.1 ± 0.8 × $10^3\ M^{-1}\ s^{-1}$. In order to minimize the possibility that the authentic substrate was overlooked during mass spectroscopic screening, additional L-Ala-L-Xxx dipeptides were characterized. These dipeptides were specifically chosen to systematically sample the different classes of amino acid side chains in the second position, regardless of apparent turnover in MS assays. We found that L-Ala-L-Glu and L-Ala-L-Leu were epimerized with values of k_{cat}/K_M of 4.9 ± 1 × $10^3\ M^{-1}\ s^{-1}$ and 3.8 ± 1 × $10^3\ M^{-1}\ s^{-1}$, respectively. These results indicate that, although not optimal, negative and aliphatic side chains can also be accommodated in the C-terminal position. Finally, low turnover of L-Ala-L-Lys, 3.6 ± 0.2 × $10^2\ M^{-1}\ s^{-1}$, indicates that a positively charged group in the epimerized position is detrimental.

Kinetic constants were also determined for selected compounds from the L-Xxx-L-Phe and L-Xxx-L-His series. Although most of the dipeptides analyzed could serve as substrates, none had kinetic constants that approached the values of k_{cat}/K_M of $10^4\ M^{-1}\ s^{-1}$ observed for dipeptides with L-Ala in the first position. Some compounds such as L-Phe-L-Phe exhibited low values of k_{cat} (0.21 s^{-1}), whereas others such as L-Lys-L-Phe and L-Ile-L-Phe had high values of K_M. No detectable activity was observed for epimerization of L-Asp-L-Phe. Although L-Ala-L-His was a favored substrate with the value of k_{cat}/K_M essentially the same as that for L-Ala-L-Phe, other L-Xxx-L-His dipeptides were problematic substrates, with either no activity, inability to reach saturation, or evidence of substrate inhibition (Table 3). Taken together, the results support L-Ala as the optimal N-terminal residue.

Although the kinetic parameters determined for L-Ala-L-Phe, L-Ala-L-Tyr, and L-Ala-L-His at room temperature are in the range we expected for an authentic dipeptide epimerase, *T. maritima* is a hyperthermophile whose optimal growth occurs at 80°C. Although we were unable to perform the assays at temperatures elevated to this level, we were able to examine epimerization of L-Ala-L-Phe at 40°C and 50°C; the values of k_{cat}/K_M were found to be 4.1 ± 0.9 × $10^4\ M^{-1}\ s^{-1}$ and 5.4 ± 0.4 × $10^4\ M^{-1}\ s^{-1}$ at 40°C and 50°C, respectively. The values of k_{cat} double with each 10°C increase (from 16 ± 7 s^{-1} at 28°C to 35 ± 6 s^{-1} at 40°C, and to 76 ± 20 s^{-1} at 50°C). From these results we conclude that the measured kinetic parameters likely underestimate the physiological efficiency of the enzyme. The physiologically relevant substrate is currently unknown, but we consider L-Ala-L-Phe, L-Ala-L-Tyr, and L-Ala-L-His to be the most likely candidates based on their kinetic constants.

The homology model revealed the structural basis for the change in specificity (Figure 2). One critical determinant of specificity in the *B. subtilis* and closely related AEEs is Arg24, which

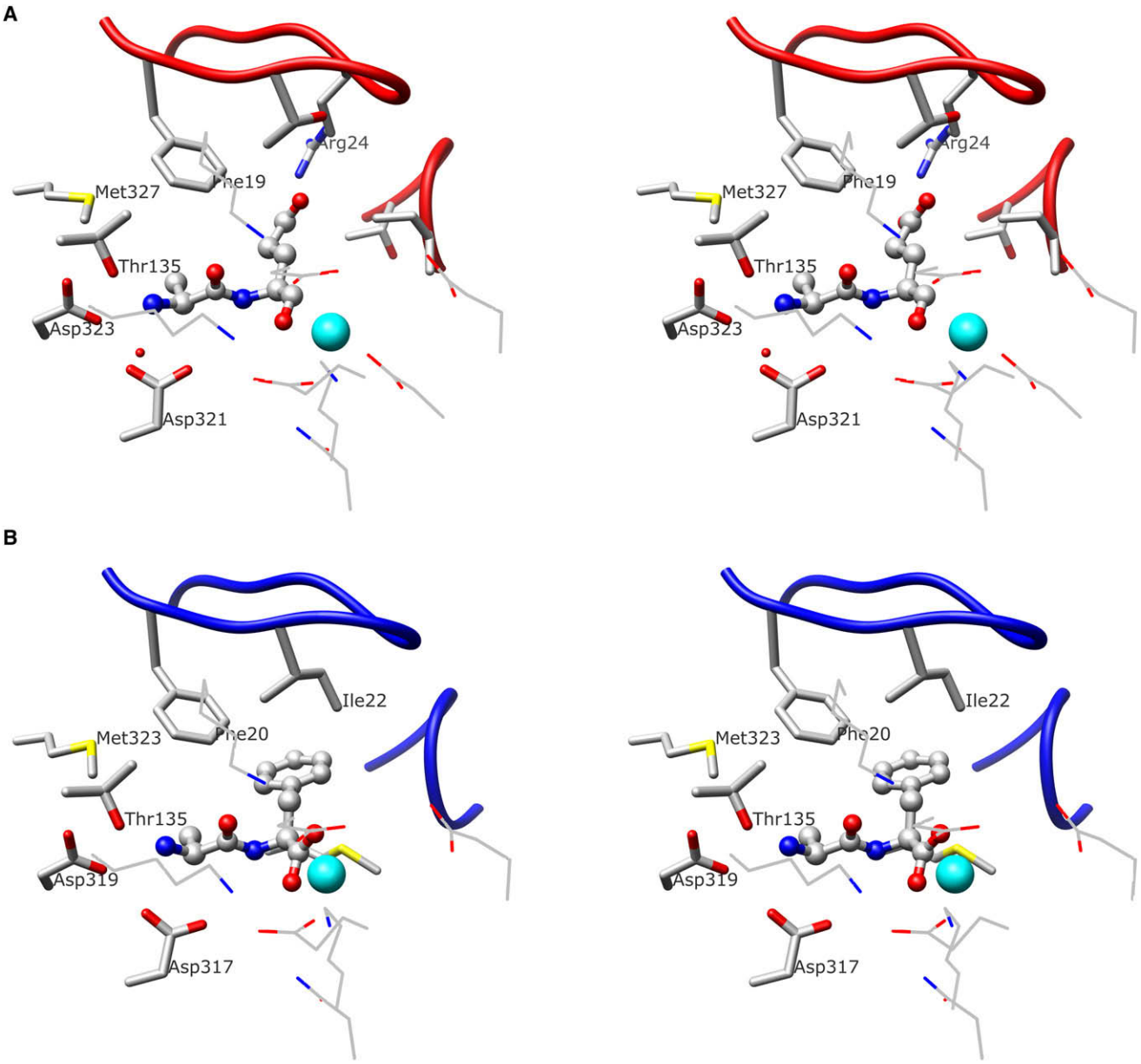

Figure 2. Stereo View Depictions of the Dipeptide-Binding Site in the *B. subtilis* AEE and TM0006

(A and B) The dipeptide-binding site in the *B. subtilis* AEE is shown at the top of the figure (PDB ID: 1TKK, cocrystallized with L-Ala-L-Glu), and that for TM0006 is shown at the bottom (homology model, with docked L-Ala-L-Phe). The dipeptides are in ball-and-stick representation. Key specificity determinants are highlighted and labeled. The Mg^{2+} ion is a blue sphere. Catalytic and metal-binding side chains are represented by thin lines. The backbone of the 20 s and 50 s loops are shown as tubes.

coordinates the Glu side chain of the L-Ala-L-Glu ligand. The corresponding residue in TM0006 is Ser25 (Figure 3). Other members of the dipeptide epimerase group also have substitutions at this position, including the *E. coli* AEE, which has Gly24 at the equivalent position. The specificity for Glu in *E. coli* AEE and related proteins is provided by Arg and Lys side chains at other positions within the same pocket. The pocket in TM0006, however, is primarily hydrophobic, accounting for the change in specificity. With respect to the N-terminal position of the substrate, the ability to accommodate side chains larger than Ala/Ser/Thr is conferred in part by the substitution of Gly294 at the position equivalent to Ile298 in the *B. subtilis* AEE.

The crystal structure of TM0006 was subsequently determined as an apo structure as well as in complex with L-Ala-L-Phe, L-Ala-L-Leu, and L-Ala-L-Lys, at 1.9–2.3 Å resolution (Figure 4). During the preparation of this manuscript, an apo structure for an ortholog of TM0006 was released in the PDB (2ZAD; currently unpublished). This structure was not available when this work was performed, and it agrees closely with the apo structure determined here. The experimentally determined structure of the

A

```
                  1                                                    60
15642781_thema    MSRIVNVKLS LKRYEYEKPF HITGSVSSES RNVEVEIVLE SGVKGYGEAS PSFRVNGERV
82500522_calsa    MSKIVNVKFD VKQYKYEKPF HITGSISSET KNVEVEVILE SGAKGCGEAA PSFRVNGEKI
55229873_halma    MTRIERVSAE PLEHELDEPF EISLGTREKA RNLVVAVETD SGIVGHGEGS PLPFVTGETQ
22971600_chlau    STVIRTLSVA PLDIPLRSPF GIAGGAQEVA RNLLVTVELM NGVRGYGEAA PFPAFNGETQ
55770159_orysa    SVDVVAAEAR PLNVPLAAPF TIASSRLDAV SNVAVRVELR SGAVGWGEAP VLPSVTAEDQ
18401824_arath    TVRVLKAENR ELNVALLSPF TIASSRLDSV SNVAIRIELN DGFVGWGEAP ILPSVTAEDQ
16078363_bacsu    .MKIIRIETS RIAVPLTKPF KTALRTVYTA ESVIVRITYD SGAVGWGEAP PTLVITGDSM
2506874_escco     ...MRTVKVF EEAWPLHTPF VIARGSRSEA RVVVVELEEE .GIKGTGECT PYP.RYGESD
```

B

```
                  301                                                  360
15642781_thema    SGLKLMIGCM GESSLGINQS VHFALGTGAF EFHDLDSHLM LKEEVFRGKF IQDGPRMRVK
82500522_calsa    ANLRLMIGCM SESSVGINQS VHFALGTGAF DFHDLDSHLM LQEPGFRGKF IQSGAKMIGY
55229873_halma    ASLDCMLGCM LEGATGIATS AHLAAGLGAF DYVDLDGHLL FEECP...PS MSSGPRIDI.
22971600_chlau    AGLGLMIGGM VESILAMTVS AHFAAGLGGF RFVDLDTPLF LATNPFQGGF ELRGGVIDLC
55770159_orysa    ARIALMIGGM VETRIAMGFA GHLAAGLGCF SFVDLDTPLL LSEDPVFGGY EVSGPVYKFT
18401824_arath    SGIELMIGGM VETRLAMGFS GHLAAGLGCF RFIDLDTPLL LADDPVQGGY KACGAVYEFK
16078363_bacsu    CGVECMVGSM IETKLGITAA AHFAASKRNI TRFDFDAPLM LKTDVFNGGI TYSGSTISMP
2506874_escco     QGFSLMLGCM LCTSRAISAA LPLV..PQ.V SFADLDGPTW LAVDV.EPAL QFTTGELHL.
```

C

B. subtilis AEE	Phe19	Thr21	Arg24	Thr51	Ile54	Ile298	Asp321	Phe322	Asp323	Ala324
TM0006	Phe20	Ile22	Ser25	Ser52	Val55	Gly294	Asp317	Leu318	Asp319	Ser320
	C-terminal side chain pocket					N-terminal residue (NH_3^+ and side chain)				

Figure 3. Portions of the Multiple Sequence Alignment of TM0006, Several of Its Closest Homologs Based on the Phylogenetic Tree, and the *E. coli* and *B. subtilis* AEEs

(A–C) Residues in TM0006 that directly contact the side chain or NH_3^+ terminus of the dipeptide substrates are highlighted; these are also highlighted in the other sequences if they are conserved. Residues involved in catalysis and metal binding are found in other portions of the sequence alignment and are not shown here (see Supplemental Data). (A) The N-terminal portion of the alignment, which includes the residues involved in binding the side chain of the epimerized residue in the dipeptide ligand. (B) The C-terminal portion of the alignment, which includes the residues contacting the N-terminal residue of the dipeptide ligand. (C) Corresponding specificity-determining residues in the *B. subtilis* AEE and TM0006 sequences (27% sequence identity overall).

L-Ala-L-Phe complex is superimposed on the model generated by homology modeling and docking in Figure 5. The experimental structure confirmed the proposed binding mode; the ligands superimpose almost perfectly. The positions of most of the protein side chains in the immediate vicinity of the ligand were also predicted accurately, reflecting no major errors in the sequence alignment used to generate the homology model. The greatest discrepancy is between the predicted and observed position of Arg54, which forms a salt-bridging interaction with Glu242 in the crystal structure. In the computational model, Arg54 is swung out into solution. This error may be due to a slight shift in the backbone near Arg54 between the homology model and the

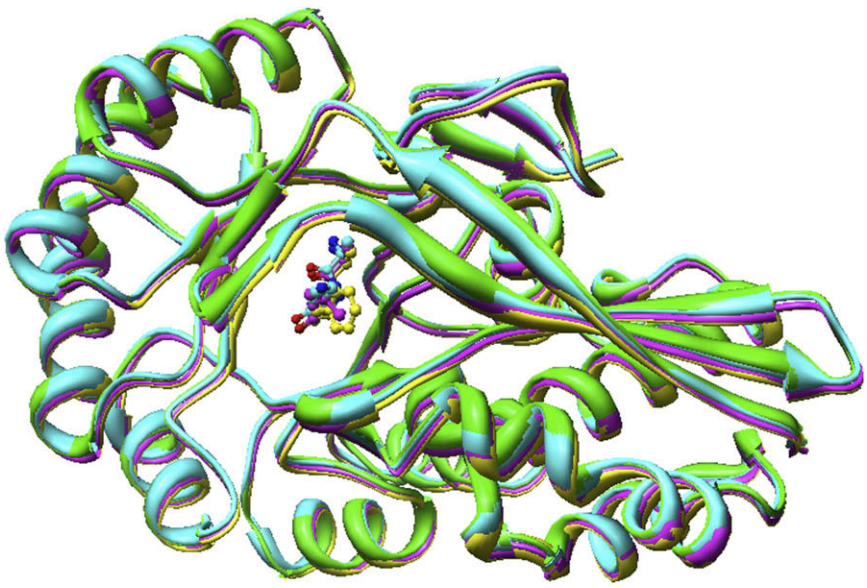

Figure 4. Overview of Structures of TM0006 Obtained by X-Ray Crystallography

Only chain A from each structure is depicted. The dipeptide ligands in the holo structures are shown in ball-and-stick representation, to define the active site, which is located in the α-β barrel domain, and capped by the "20s" loop from the N-terminal domain. Green, apo structure (3DFY); magenta, complex with L-Ala-L-Leu (3DEQ); cyan, complex with L-Ala-L-Lys (3DER); yellow, complex with L-Ala-L-Phe (3DES).

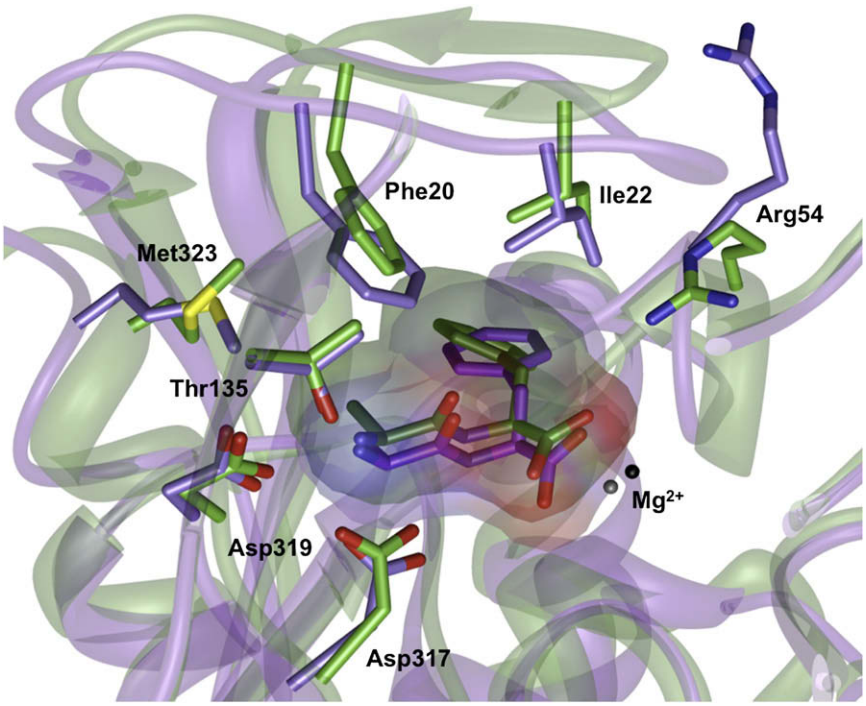

Figure 5. Superposition of the Models of L-Ala-L-Phe Bound to TM0006, Based on Homology Modeling and Docking and Crystallography

The computational model is shown in purple, and crystal structure is shown in green.

crystal structure, to a limitation of the energy function used for constructing the homology model, or both. Arg54 may play some role in substrate specificity, because it comes within 4 Å of the Phe side chain of the dipeptide ligand in the crystal structure, possibly forming a favorable cation-pi interaction.

The active sites of the other holo structures are shown in Supplemental Data (available online). The complex with L-Ala-L-Lys was determined to elucidate the structural basis for the relatively slow but detectable epimerization for this dipeptide, which is positively charged, in contrast to most of the other substrates, which are hydrophobic. The structure of the complex of TM0006 with L-Ala-L-Lys reveals that the positively charged nitrogen of the Lys side chain extends slightly out of the binding pocket through a narrow opening and is coordinated by water molecules.

DISCUSSION

We know of no other enzymes that epimerize hydrophobic dipeptides. The dipeptide epimerase from *T. maritima* clusters with a few other sequences that we also predict not to be AEEs, based on the sequence analysis, homology models, and docking results (Figures 1 and 3). So far, it has not been possible to express these proteins in soluble form for experimental screening. The other members of this small group include proteins from other thermophiles, *Caldicellulosiruptor saccharolyticus* and *Chloroflexus aurantiacus*, and an archaeon, *Haloarcula marismortui*. Most strikingly, the group also contains sequences from two plant genomes, *Oryza sativa* (rice) and *Arabidopsis thaliana*. Although the physiological relevance is not clear, the presence of closely related dipeptide epimerases in organisms lacking peptidoglycan is consistent with the change in specificity that we have described.

This study highlights the challenges facing functional assignment of enzymes, as well as a promising approach for overcoming some of these challenges. A central challenge is delineating, in sequence space, where one function ends and another begins. Overall sequence similarity among proteins is often unreliable, especially in mechanistically diverse superfamilies (Pegg et al., 2006). Changes in enzymatic function are often related to sequence changes in the binding site, and, as shown here, homology models can be used to identify proteins that are likely to have different functions than their closest functionally and structurally characterized homologs.

Docking methods can be used in conjunction with homology models to suggest specific small molecules that may be substrates, and, importantly, can suggest novel enzymatic functions, as we, to our knowledge, have done here. At this point, experimental testing remains necessary to confirm or refute these hypotheses. Predicting catalytic rates remains extremely difficult, and we have not attempted to do so. Predicting substrates likely to bind, as well as their binding modes, is more tractable, although still challenging, especially with homology models (Jacobson and Sali, 2004). In this case, most of the top dipeptides from docking were in fact substrates. Although the top docking hits were not necessarily the best substrates (L-Ala-L-Phe ranked 48 out of 400 dipeptides), they did capture the correct specificity at the epimerized position for aromatic side chains. Most importantly, the striking differences between the docking hit lists for TM0006 and the *B. subtilis* and *E. coli* AEEs allowed us to identify TM0006 as a candidate for experimental screening based on the high likelihood of it epimerizing distinct substrates.

We believe that the integrated use of computational methods (multiple sequence alignment, operon context, phylogenetic trees, homology modeling, and docking) applied on the scale of hundreds or thousands of proteins, in combination with experimental characterization (functional enzymology and structural biology) of a relatively small number of proteins that are predicted to have new functions, will be a powerful approach for accurate and large-scale functional annotation.

EXPERIMENTAL PROCEDURES

Computational Methods

All protein sequences annotated in the Structure Function Linkage Database (Pegg et al., 2005, 2006) as belonging to the muconate lactonizing enzyme (MLE) subgroup were used to construct the multiple sequence alignment.

The proteins were first aligned by using Muscle v.3.52 (Edgar, 2004), and the initial alignment was manually refined by referring to structural alignments of the characterized MLE subgroup members (Glasner et al., 2006). The phylogenetic tree was constructed by using MrBayes v3.1.2 (Altekar et al., 2004; Ronquist and Huelsenbeck, 2003) under the WAG amino acid substitution model (Whelan and Goldman, 2001) and a gamma distribution to approximate the rate variation among sites. Positions in the alignment that had too many gaps or appeared to be mutationally saturated were excluded from phylogenetic analysis. Accession numbers and species abbreviations are listed in Table S1.

Homology models were created for over 100 sequences in the MLE subgroup, including 82 sequences that clustered with the experimentally characterized AEEs from *B. subtilis* and *E. coli* according to the phylogeny (Glasner et al., 2006). At the time of our investigation, there were only three crystal structures available among the sequences in this clade: holo (cocrystallized with L-Ala-L-Glu) and apo structures of *B. subtilis* AEE and an apo structure of the *E. coli* AEE. We used the holo structure of the *B. subtilis* AEE (1TKK) as a template to construct models for the 82 sequences. The models were built by using the Protein Local Optimization software (marketed as Prime by Schrödinger LLC). While constructing the models, we included both the metal ion and the cocrystallized ligand from the template. After building the models, we docked a dipeptide library against the binding site of these models by using the software Glide (v4.0108, Schrödinger LLC). The dipeptide library was prepared by using the software Ligprep (v2.0106, Schrödinger LLC).

Cloning, Expression, and Purification of the Dipeptide Epimerase from *Thermotoga maritima*

The gene for the dipeptide epimerase (gi:15642781) was amplified by PCR from *Thermotoga maritima* MSB8 genomic DNA by using the following primers: 5′-GGAGGTGTGACATATGTCGAGGATCGTGAACGTGAAGC-3′ and 5′-GAACTGCTGGATCCTCATTGATCTTTCACCCTCATTCTCG-3′ (Bio-Synthesis, Inc.) containing a 5′ NdeI site and a 3′ BamHI site, respectively. PCR reactions in 100 μl total volume contained 1 ng template, 1 mM $MgSO_4$, 2.5 U platinum *Pfx* DNA polymerase (Invitrogen), 1× *Pfx* amplification buffer, 1× enhancer buffer, 0.4 mM of each dNTP, and 0.2 μM of each forward and reverse primer. The PCR reaction was performed with the following parameters: 94°C for 3 min, followed by 40 cycles of 94°C for 1 min, 47°C for 1.25 min, and 68°C for 3 min; the final extension time was 10 min at 68°C. After purification by gel extraction (QIAGEN), the amplified PCR product was restricted by using NdeI and BamHI restriction enzymes (New England Biolabs) per the manufacturer's protocols. The gene was then ligated into the nontagged expression vector pET17b (Novagen) by using T4 DNA ligase (Fisher) and was transformed in *E. coli* XL1Blue cells for plasmid amplification and maintenance.

The cloned dipeptide epimerase from *T. maritima* was expressed in *E. coli* BL21 (DE3) cells for protein purification. In a typical protein preparation, 2 L LB media was shaken at 37°C without induction and harvested after 32 hr by centrifugation at 4800 rpm. The pelleted cells were resuspended in 60 ml buffer containing 10 mM Tris-HCl (pH 7.9) and 5 mM $MgCl_2$. The suspension was lysed by sonication, and debris was cleared by centrifugation at 27,250 × g. The supernatant was applied to a DEAE Sepharose FF column (2.5 × 50 cm, GE Healthcare) and eluted with a linear gradient (1600 ml) of 0 to 1 M NaCl buffered with 10 mM Tris-HCl (pH 7.9) containing 5 mM $MgCl_2$. Fractions containing the protein of interest were pooled and dialyzed three times against 10 mM Tris-HCl (pH 7.9) containing 5 mM $MgCl_2$ before being applied to a Q Sepharose HP column (1.7 × 7 cm, GE Healthcare). The protein was eluted with a linear gradient (250 ml) of 0 to 0.5 M NaCl in 10 mM Tris-HCl (pH 7.9) containing 5 mM $MgCl_2$. Fractions containing >99% pure protein were pooled and dialyzed into 20 mM Tris-HCl (pH 7.9) containing 100 mM NaCl and 5 mM $MgCl_2$. The protein was concentrated to 10–15 mg/ml by using a Millipore Amicon apparatus fitted with a 10,000 NMWL ultrafiltration membrane and was stored at 4°C. Storage for more than 1 week resulted in an ~25% loss of activity.

Repeated attempts to achieve expression in an *E. coli* AEE knockout system failed. As an alternative, endogenous *E. coli* proteins were heat denatured during purification. After initial sonication and centrifugation, the cleared lysate (vide supra) was heated at 50°C for 60 min until the solution was opaque and viscous. Centrifugation at 27,250 × g for 50 min was repeated, and the lysate was further purified over DEAE/Q Sepharose columns as described above. This preparation was used to assess the validity of the AEE activity of the *Thermotoga* enzyme, which is elaborated on in the Supplemental Data. The *E. coli* AEE was purified as previously reported (Schmidt et al., 2001).

Screening of the *Thermotoga* Enzyme with Dipeptide Libraries

The procedure for solid-state synthesis of dipeptide libraries was reported previously (Song et al., 2007). For initial assessment of dipeptide epimerase activity, screens were set up with the following dipeptide libraries: Gly-L-Xxx, L-Ala-L-Xxx, D-Ala-L-Xxx, L-Thr-L-Xxx, L-Xxx-L-Phe, L-Xxx-L-Tyr, L-Xxx-L-His, and L-Xxx-L-Lys. In accordance with known activities in the MLE subgroup of the enolase superfamily, screens with *N*-succinyl-L-Xxx and *N*-acetyl-L-Xxx libraries were also preformed. Screens were carried out in 50 μl D_2O containing 20 mM NH_4HCO_3 (pD 7.9), 1 mM (each dipeptide) library, and 1 μM enzyme. The reaction was incubated at 37°C for 16 hr, quenched with 2 μl 5 M NH_4OH, and evaporated to dryness. The samples were then resuspended in ddH_2O and analyzed by ESI mass spectrometry for incorporation of solvent deuterium as indicated by a +1 mass shift. If activity was detected, the screen was repeated with a 10-fold reduction in enzyme, and time points were taken at 0.5, 1.5, and 3 hr for better assessment of preferred substrates.

Kinetic Studies of the *Thermotoga* Enzyme with Dipeptide Substrates

Polarimetry measurements were determined on a Jasco P-1010 Polarimeter. Dipeptides for kinetic characterization were purchased when possible (e.g., Sigma, Bachem, Research Organics, Indofine, or MP Biochemicals), with the following exceptions: L-Ile-L-Phe was synthesized according to the procedure of Theodoropoulos and Craig (1955), and L-Lys-L-Phe was synthesized according to that of Lapeyre et al. (2006). Syntheses for all other dipeptides are provided in Supplemental Data. Kinetic parameters were obtained by quantifying the change in optical rotation as a function of time as determined by polarimetry by using a 100 mm path cell and an Hg 405 nm filter. Assays were performed at room temperature (~28°C) in 1.4 ml total volume containing 20 mM Tris-Cl (pH 7.5) and 10 mM $MgCl_2$, with variable enzyme and substrate concentrations. The molar ellipticities for epimerized dipeptides were determined by subtracting the optical rotation at equilibrium from the starting optical rotation. Rate constants were divided by two to account for reversibility. Values for k_{cat} and K_M were determined by fitting initial velocities to Michealis-Menton curves by using the program EnzFitter (Madison, WI). Errors presented are standard deviations determined from a minimum of three independent sets of kinetic assays. Kinetic parameters determined at 40°C and 50°C were quantified as described above, by using a 100 mm path-length water-jacketed cell connected to a Fisher Isotemp water bath (model 9000). Temperatures were monitored via a sensor in direct contact with the reaction solution.

Crystallization and Data Collection

Four different crystal forms (Table 4) were grown by the hanging-drop method at room temperature: (1) TM0006 in complex with Mg^{2+}, (2) TM0006 in complex with Mg^{2+} and L-Ala-L-Leu, (3) TM0006 in complex with Mg^{2+} and L-Ala-L-Lys, and (4) TM0006 in complex with Mg^{2+} and L-Ala-L-Phe. The crystallization conditions were as follows:

(1) For TM0006 in complex with Mg^{2+}, the protein solution contained TM0006 (22.3 mg/ml) in 20 mM Tris (pH 7.9), 100 mM NaCl, and 10 mM $MgCl_2$; the precipitant contained 10% PEG 6000, 0.1 M HEPES (pH 7.5), and 5% MPD. For this sample, crystals appeared in 7–8 days and exhibited diffraction consistent with the space group $P2_1$, with 16 molecules of TM0006 per asymmetric unit.

(2) For TM0006 in complex with Mg^{2+} and L-Ala-L-Leu, the protein solution contained TM0006 (33 mg/ml) in 20 mM Tris (pH 7.9), 100 mm NaCl, 10 mM $MgCl_2$, and 20 mM L-Ala-L-Leu; the precipitant contained 3.6 M NaCl and 0.1 M CH_3COONa (pH 4.5). For this and the remaining samples, crystals appeared in 2 days and exhibited a diffraction pattern consistent with space group $P6_122$, with four molecules of dipeptide epimerase per asymmetric unit.

(3) For TM0006 in complex with Mg^{2+} and L-Ala-L-Lys, the protein solution contained TM0006 (33 mg/ml) in 20 mM Tris (pH 7.9), 100 mM NaCl,

Table 4. X-Ray Data Collection and Refinement Statistics for Crystals of TM0006

	Complex with Mg^{2+}	Complex with Mg^{2+} and Ala-Leu	Complex with Mg^{2+} and Ala-Lys	Complex with Mg^{2+} and Ala-Phe
Data Collection				
Wavelength (Å)	0.979	0.979	0.979	0.979
Space group	$P2_1$	$P6_122$	$P6_122$	$P6_122$
Mol. in a.u.	16	4	4	4
Unit cell parameters				
a (Å)	104.81	191.58	190.69	191.67
b (Å)	165.14			
c (Å)	209.64	283.23	283.11	283.08
β (°)	96.06			
Resolution (Å)	25.0–2.1	25.0–2.1	25.0–1.9	25.0–2.3
Unique reflections	407,187	172,237	230,926	135,031
Completeness (%)	98.9	97.5	97.8	99.7
R_{merge}	0.064	0.092	0.084	0.057
Average I/σ	17.4	23.6	25.9	21.6
Refinement				
Resolution (Å)	25.0–2.1	25.0–2.1	25.0–1.9	25.0–2.3
R_{cryst}	0.246	0.245	0.231	0.213
R_{free}	0.277	0.259	0.244	0.231
Rmsd, bonds (Å)	0.006	0.006	0.005	0.006
Rmsd, angles (°)	1.30	1.33	1.29	1.20
Number of atoms				
Protein	42,058	10,755	10,755	10,755
Water	1072	386	719	446
Mg^{2+}	16	4	4	4
Bound peptide		56	60	68
PDB entry	3DFY	3DEQ	3DER	3DES

10 mM $MgCl_2$, and 40 mM L-Ala-L-Lys; the precipitant contained 3.4 M NaCl and 0.1 M CH_3COONa (pH 4.5).

(4) For TM0006 in complex with Mg^{2+} and L-Ala-L-Phe, the protein solution contained TM0006 (33 mg/ml) in 20 mM Tris (pH 7.9), 100 mM NaCl, 10 mM $MgCl_2$, and 60 mM L-Ala-L-Phe; the precipitant contained 3.0 M NaCl and 0.1 M CH_3COONa (pH 4.5).

Prior to data collection, the crystals were transferred to cryoprotectant solution composed of their mother liquids and 20% glycerol and were flash cooled in a nitrogen stream. All X-ray diffraction data sets for the complexes of TM0006 with Mg^{2+} (Table 4, column 1), with Mg^{2+} and L-Ala-L-Leu (column 2), with Mg^{2+} and L-Ala-L-Lys (column 3), and with Mg^{2+} and L-Ala-L-Phe (column 4) were collected at the NSLS X4A beamline (Brookhaven National Laboratory) on an ADSC CCD detector to 2.1, 2.1, 1.9, and 2.3 Å resolution, respectively. Diffraction intensities were integrated and scaled with DENZO and SCALEPACK, respectively (Otwinowski and Minor, 1997). The data collection statistics are given in Table 4.

Structure Determination and Model Refinement

The structure of apo TM0006 was solved by molecular replacement with the fully automated molecular replacement pipeline BALBES (Long et al., 2008), by using only input diffraction and sequence data. The partially refined structure of apo TM0006 was output from BALBES without any manual intervention. Subsequently, several iterative cycles of manual rebuilding with TOM (Jones, 1985), refinement with CNS (Brunger et al., 1998), and automatic rebuilding with ARP (Lamzin and Wilson, 1993) resulted in a model with an R_{cryst} and R_{free} of 0.246 and 0.277, respectively. The final structure contains 42,058 protein atoms, 1,072 water molecules, and 16 Mg^{2+} ions for 2 octamers of TM0006 in the asymmetric unit. Both TM0006 octamers are similar to the octamers observed in the *E. coli* and *B. subtilis* AEE epimerases (PDB files 1JPD and 1JPM, respectively). For the apo TM0006 structure (Table 4, column 1) and for the holo TM0006 structures (columns 2, 3, and 4), no nonglycine residues lie in the disallowed region of the Ramachandran plot. Residues 325–327 have no density in 4 out of 16 monomers and are not included in the final model. Also, flap regions 19–27 are not included in the final model for 12 monomers out of 16. The Mg^{2+} ions are well defined in all 16 monomers in the asymmetric unit. Each Mg^{2+} ion is coordinated by the side chains of Asp188, Glu216, and Asp241 and by three water molecules in each TM0006 monomer.

The structure of TM0006 crystallized with Mg^{2+} and L-Ala-L-Leu was automatically solved and partially refined with BALBES by using corresponding X-ray and sequence data. Subsequent iterative cycles of manual rebuilding with TOM, refinement with CNS, and automatic rebuilding with ARP were performed. The model was refined at 2.1 Å with an R_{cryst} of 0.245 and an R_{free} of 0.259. The final structure contained residues (3–343), Mg^{2+} ions, and bound dipeptide with well-defined density in all four monomers of the asymmetric unit. The Mg^{2+} ion is coordinated by side chains of Asp188, Glu216, and Asp241; by one water molecule; and by two oxygen atoms from the dipeptide carboxyl terminus.

The protein portion of the complex with Mg^{2+} and L-Ala-L-Leu was the starting point for the refinement of TM0006 crystallized with Mg^{2+} and L-Ala-L-Lys (Table 4, column 3) and TM0006 crystallized with Mg^{2+} and L-Ala-L-Phe (column 4). These three structures contain the same protein molecules crystallized in the same space group. Iterative cycles of manual rebuilding with TOM, refinement with CNS, and automatic rebuilding with ARP with subsequent inclusion of water molecules were performed for the complexes with L-Ala-L-Lys and with L-Ala-L-Phe. Mg^{2+} ions were clearly defined in each monomer of both complexes and have the coordination identical to that found in the complex with L-Ala-L-Leu.

Final crystallographic refinement statistics are provided in Table 4.

ACCESSION NUMBERS

Coordinates have been deposited in the PDB with accession codes 3DFY, 3DEQ, 3DER, and 3DES.

SUPPLEMENTAL DATA

Supplemental Data include one table, two figures, and Supplemental Methods and can be found with this article online at http://www.structure.org/cgi/content/full/16/11/1668/DC1/.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grant P01-GM071790. M.P.J. is a consultant to Schrödinger, Inc.

Received: June 24, 2008
Revised: August 14, 2008
Accepted: August 19, 2008
Published: November 11, 2008

REFERENCES

Altekar, G., Dwarkadas, S., Huelsenbeck, J.P., and Ronquist, F. (2004). Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics *20*, 407–415.

Barker, J.A., and Thornton, J.M. (2003). An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. Bioinformatics *19*, 1644–1649.

Broun, P., Shanklin, J., Whittle, E., and Somerville, C. (1998). Catalytic plasticity of fatty acid modification enzymes underlying chemical diversity of plant lipids. Science *282*, 1315–1317.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. *54*, 905–921.

Cammer, S.A., Hoffman, B.T., Speir, J.A., Canady, M.A., Nelson, M.R., Knutson, S., Gallina, M., Baxter, S.M., and Fetrow, J.S. (2003). Structure-based active site profiles for genome analysis and functional family subclassification. J. Mol. Biol. *334*, 387–401.

de Souza, M.L., Seffernick, J., Martinez, B., Sadowsky, M.J., and Wackett, L.P. (1998). The atrazine catabolism genes atzABC are widespread and highly conserved. J. Bacteriol. *180*, 1951–1954.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Favia, A.D., Nobeli, I., Glaser, F., and Thornton, J.M. (2008). Molecular docking for substrate identification: the short-chain dehydrogenases/reductases. J. Mol. Biol. *375*, 855–874.

Glasner, M.E., Fayazmanesh, N., Chiang, R.A., Sakai, A., Jacobson, M.P., Gerlt, J.A., and Babbitt, P.C. (2006). Evolution of structure and function in the o-succinylbenzoate synthase/N-acylamino acid racemase family of the enolase superfamily. J. Mol. Biol. *360*, 228–250.

Hermann, J.C., Ghanem, E., Li, Y., Raushel, F.M., Irwin, J.J., and Shoichet, B.K. (2006). Predicting substrates by docking high-energy intermediates to enzyme structures. J. Am. Chem. Soc. *128*, 15882–15891.

Hermann, J.C., Marti-Arbona, R., Fedorov, A.A., Fedorov, E., Almo, S.C., Shoichet, B.K., and Raushel, F.M. (2007). Structure-based activity prediction for an enzyme of unknown function. Nature *448*, 775–779.

Jacobson, M., and Sali, A. (2004). Comparative protein structure modeling and its applications to drug discovery. Annu. Rep. Med. Chem. *39*, 259–276.

Jones, T.A. (1985). Interactive computer-graphics: Frodo. Methods Enzymol. *115*, 157–171.

Kalyanaraman, C., Bernacki, K., and Jacobson, M.P. (2005). Virtual screening against highly charged active sites: identifying substrates of α-β barrel enzymes. Biochemistry *44*, 2059–2071.

Kenyon, V., Chorny, I., Carvajal, W.J., Holman, T.R., and Jacobson, M.P. (2006). Novel human lipoxygenase inhibitors discovered using virtual screening with homology models. J. Med. Chem. *49*, 1356–1363.

Klenchin, V.A., Schmidt, D.M., Gerlt, J.A., and Rayment, I. (2004). Evolution of enzymatic activities in the enolase superfamily: structure of a substrate-liganded complex of the L-Ala-D/L-Glu epimerase from *Bacillus subtilis*. Biochemistry *43*, 10370–10378.

Lamzin, V.S., and Wilson, K.S. (1993). Automated refinement of protein models. Acta Crystallogr. D Biol. Crystallogr. *49*, 129–147.

Lapeyre, M., Leprince, J., Massonneau, M., Oulyadi, H., Renard, P.-Y., Romieu, A., Turcatti, G., and Vaudry, H. (2006). Aryldithioethyloxycarbonyl (Ardec): a new family of amine protecting groups removable under mild reducing conditions and their applications to peptide synthesis. Chem. Eur. J. *12*, 3655–3671.

Long, F., Vagin, A.A., Young, P., and Murshudov, G.N. (2008). BALBES: a molecular-replacement pipeline. Acta Crystallogr. D Biol. Crystallogr. *64*, 125–132.

Macchiarulo, A., Nobeli, I., and Thornton, J.M. (2004). Ligand selectivity and competition between enzymes in silico. Nat. Biotechnol. *22*, 1039–1045.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. In Methods in Enzymology, C.W.J. Carter, R.M. Sweet, J.N. Abelson, and M.I. Simon, eds. (New York: Academic Press), pp. 307–326.

Pegg, S.C., Brown, S., Ojha, S., Huang, C.C., Ferrin, T.E., and Babbitt, P.C. (2005). Representing structure-function relationships in mechanistically diverse enzyme superfamilies. Pac. Symp. Biocomput. *10*, 358–369.

Pegg, S.C., Brown, S.D., Ojha, S., Seffernick, J., Meng, E.C., Morris, J.H., Chang, P.J., Huang, C.C., Ferrin, T.E., and Babbitt, P.C. (2006). Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. Biochemistry *45*, 2545–2555.

Polacco, B.J., and Babbitt, P.C. (2006). Automated discovery of 3D motifs for protein function annotation. Bioinformatics *22*, 723–730.

Ronquist, F., and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics *19*, 1572–1574.

Schmidt, D.M., Hubbard, B.K., and Gerlt, J.A. (2001). Evolution of enzymatic activities in the enolase superfamily: functional assignment of unknown proteins in *Bacillus subtilis* and *Escherichia coli* as L-Ala-D/L-Glu epimerases. Biochemistry *40*, 15707–15715.

Seffernick, J.L., de Souza, M.L., Sadowsky, M.J., and Wackett, L.P. (2001). Melamine deaminase and atrazine chlorohydrolase: 98 percent identical but functionally different. J. Bacteriol. *183*, 2405–2410.

Song, L., Kalyanaraman, C., Fedorov, A.A., Fedorov, E.V., Glasner, M.E., Brown, S., Imker, H.J., Babbitt, P.C., Almo, S.C., Jacobson, M.P., and Gerlt, J.A. (2007). Prediction and assignment of function for a divergent N-succinyl amino acid racemase. Nat. Chem. Biol. *3*, 486–491.

Theodoropoulos, D., and Craig, L.C. (1955). The synthesis of several isoleucyl peptides and certain of their property. J. Org. Chem. *20*, 1169–1172.

Tian, W., and Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? J. Mol. Biol. *333*, 863–882.

Tremblay, L.W., Dunaway-Mariano, D., and Allen, K.N. (2006). Structure and activity analyses of *Escherichia coli* K-12 NagD provide insight into the evolution of biochemical function in the haloalkanoic acid dehalogenase superfamily. Biochemistry *45*, 1183–1193.

Wangikar, P.P., Tendulkar, A.V., Ramya, S., Mali, D.N., and Sarawagi, S. (2003). Functional sites in protein families uncovered via an objective and automated graph theoretic approach. J. Mol. Biol. *326*, 955–978.

Whelan, S., and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. *18*, 691–699.